Secondary Databases

Derivative databases use data from primary database like GenBank and add value by performing some kind of computational analysis or additional annotation and curation.

Protein only databases

Currently protein only sequence databases like PIR,the Protein Information Resource (URL: http://pir.georgetown.edu/)and SWISS-PROT (URL: http://www.expasy.ch/sprot/)are essentially derivative database because the majority of protein sequences in them come from translations of nucleotide sequences. Both of these databases curate the protein sequences extensively and add additional annotations. These include comparing various examples of the protein sequences derived from primary sources. Both the SWISS-PROT data and PIR data are available at the NCBI.

NCBI Secondary Databases

UniGene

The UniGene database contains sequence similarity based clusters of expressed sequences. Naturally the richest source of expressed sequences is the EST data. These data of course over represent the number of transcripts because highly expressed messages will be represented many times within the data. The goal of the UniGene is reduce the EST data and to identify all transcripts for a particular organism. UniGene data sets are available for those organisms with substantial EST data: human, mouse, rat, zebrafish, cow and *Xenopus*. The UniGene data are a rich source for gene discovery. Because EST libraries are tissue specific, UniGene data can be used as a resource for gene expression information. The NCBI Serial Analysis of Gene expression pages and the CGAP pages take advantage of this latter feature.

UniGene Build Procedure

Expressed sequences and coding regions from genes are clustered by sequence similarity. This is done in stages after removing mitochondrial, vector sequences and masking for repetitive elements. An important problem is cross clustering of sequences for different genes. Cross clustering can arise because the level of sequencing errors in ESTs may approach the level of sequence divergence of members of the same gene family. One way to avoid some of this is to focus on the 3' untranslated regions first since these are less well conserved than coding. Then clone based edges are added, this means adding 5' reads from clones whose 3' ends have already been clustered. Although the UniGene Data sets have “value added” to the primary EST data, the databases are not truly curated but are built automatically.

LocusLink and the RefSeq Project.

Because of the tremendous growth in primary sequence data and the archival nature of these datasets, it can be difficult to identify the best sequence for a gene and in some cases even to find the sequence of interest because of confusing nomenclature problems. The LocusLink database attempts to solve some of these problems by collecting relevant links to sequences and other data in NCBI data as well as some outside databases. Each gene is assigned a stable unique identifier (Locus ID) and titles of entries are assigned based on the relevant genome nomenclature committee guidelines, the Human Genome Nomenclature Committee for the human genome. These titles are also propagated to the UniGene database as well to standardize nomenclature for the clusters. LocusLink also tracks historical name for the genes. The current scope is fruit fly, human, mouse, rat, and zebrafish.

RefSeq mRNAs and Proteins

A project that is intimately related to LocusLink is the generation of curated reference mRNA and protein sequences (RefSeqs) for the genes for the LocusLink entries. Collaborators supply information about which sequence is an appropriate representative for a gene. To generate a reference mRNA sequence, the best representative primary database sequence that has a full-length coding region is chosen. This record is used to create provisional RefSeq records. Essentially this provisional RefSeq is a copy of the of the database sequence but also includes several annotation enhancements: additional publications, aliases, LocusID number, MIM number, map information, and official gene symbol and name. These provisional RefSeqs are then subject to human review. This review process provides further enhancements to the RefSeq including extension of the using sequence data in other GenBank records, or the literature, correction sequencing errors, addition of additional publications and a summary of gene function. The final product represents a review article about the mRNA or protein. RefSeqs are availabel through LocusLink and are included in the Entrez and BLAST databases. RefSeq mRNA and protein sequences can be easily recognized by their distinctive accession numbers; NM_ followed by six digits for mRNA and NP_ followed by six digits for proteins.

Other NCBI Reference Sequences

There are several other kinds of reference sequences that are generated by projects at the NCBI.

Model Transcripts and Proteins

Closely related to the NM_ and NP_ RefSeqs are the model transcript and protein sequences. At this point these are generated only for the human genome by aligning the RefSeq mRNA to the corresponding genomic region. The genomic sequence that aligns is then used to create a model transcript and its corresponding translation. In many cases these sequences, do not match exactly the RefSeq mRNAs. This could be caused by assembly problems, sequencing error or true polymorphisms. These model sequences have distinctive accession numbers beginning with XM_ and XP_ and like the NM_ and NP_ RefSeqs are available through LocusLink, Entrez and through the BLAST databases.

NCBI Assemblies

NCBI has created it's own assembly of the human genome project data. The assembly consists of sets of contigs that are in turn built from assembling ovelapping h draft (HTG) and finished human sequence from GenBank. These large records are available thorugh LocusLink, the Entrez system and can be searched as a BLAST database on the Human Genome BLAST page. Their distinctive accession numbers begin with NT_. The chromosome records (NC_) that so far are created for the simpler complete genomes at NCBI are another RefSeq assembly.

Reference Genomic Records

The final type of RefSeq is a reference genomic record (NG_). These are created to serve as fixed regions of the human genome assembly They are needed where the sequence and placement of a region is well known but difficult or impossible to assemble automatically. An example is the beta globin cluster on chromosome 11.

A sunmary of RefSeq accessions is given below.

RefSeq Accession	Type of record
NM_, NP_	Reference mRNA, translation
XM_, XP_	Model Transcript, translation
NT_	contig
NC_	Reference Chromosome
NG_	Reference Genomic

Organism Specific Databases

A number of derivative databases exist for specific organisms. In most case website tha host these provide specialized tools and services and tehy may have pirmary data that is not available elsewhere. Selected sites are listed below.

PlasmoDB (Malaria parasite)	http://www.plasmodb.org/
Mouse Genome Informatics	http://www.informatics.jax.org/
Flybase (Fruit Fly)	http://flybase.bio.indiana.edu/
Rat Genome Database	http://rgd.mcw.edu/
Zfin (Zebrafish)	http://zfin.org/ZFIN/
Saccharomyces Genome Database	http://genome-www.stanford.edu/Saccharomyces/
The *C. elegans* genome project	http://www.sanger.ac.uk/Projects/C_elegans/

Exercises

1. Visit the SWISS-PROT site and use the quick search box to look for cystic fibrosis proteins. Use CFTR as search term. Compare these rather simple results to the mess you get if you search NCBI proteins for CFTR. Display the human CFTR record from the Swiss site. Examine the extensive annotation in the record and look at the features to find the amino acids that are glycosylated and those that are phosphorylated by protein kinase C. Also find most frequent mutation in this protein in the disease cystic fibrosis. Compare this to the simple GenBank record for the CFTR mRNA, M28668.

2. mRNA that hybridized to the EST sequence with accession number AI589465 was highly expressed in a human liver tumor sample. Use the human UniGene data to identify this gene. Link to LocusLink. What is the function of this protein? Go back to UniGene. Look at the ESTs in this cluster. How many are there? Identify a pair of ESTs that come from the same clone ID. Use BLAST two sequences to align these to the full length RefSeq mRNA. Are there any mismathches?

3. Another mRNA hybridizes to AI150058. What information can you find about this gene?

4. Use LocusLink to retrieve the entry for the MDR1 gene. What is the official gene symbol and name for this gene? What is the function of this gene? Use the HomoloGene link find corresponding UniGene cluster for this gene in rat and mouse. How many clusters seem to be related to the human gene. How can you explain this?

5. Retrieve the LocusLink entry for human BRCA1. How many splice variants are reported for this gene's transcripts? Display the RefSeq contig containing this gene. How large is it? How many GenBank records were used to construct it? How many of these are draft and how many are finished records? You may want to use the map viewer to examine the GenBank map to answer this last part.